intel

APPLICATION NOTE

AP-74

March, 1980

High Speed Memory System Design Using 2147H

Joe Altnether

Memory Components
Applications Engineering

INTRODUCTION

The Intel® 2147H is a 4096-word by 1-bit Random Access Memory, fabricated using Intel's reliable HMOS II technology. HMOS II, the second generation HMOS, is Intel's high performance n-channel silicon gate technology, making simple, high speed memory systems a reality. The purpose of this application note is to describe the 2147H operation and discuss design criteria for high speed memory systems.

TECHNOLOGY

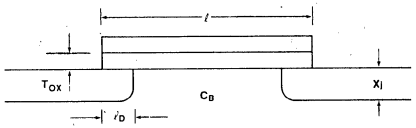
When Intel introduced the HMOS 2147, MOS static RAM performance took a quantum leap by combining scaling, internal substrate bias generation, and automatic powerdown. As a result, the 2147 has an access time of 55ns, density of 4096 bits, and power consumption of .99W active and .165W standby.

The high performance of the 2147 is further enhanced by the 2147H using HMOS II, a scaled HMOS process increasing the speed at the same power level which involves more than scaling dimensions.

Figure 1 shows the cross section of an HMOS device and lists the parameters of scaling, one of which is high device gain. The slew rate of an amplifier or device is proportional to the gain. Because faster switching speeds occur with high gain, the gain is maximized for high speed. Device gain is inversely proportional to the oxide thickness (T_{OX}) and device length (ℓ), consequently, scaling these dimensions increases the gain.

Another factor which influences performance is unwanted capacitance which appears in two forms - - diffusion and Miller. Diffusion capacitance is directly proportional to the diffusion depth (X_j) into the silicon, thus X_j must be reduced. Miller capacitance, the same phenomenon that occurs in the macro world of discrete devices, is proportional to the overlap length of the gate and the source (ℓ_D). Capacitance on the input shunts the high frequency portion of the input signal so that the device can only respond to low frequencies. Secondly, capacitance from the drain to the gate forms a feedback path creating an integrator or low pass filter which degrades the high frequency performance. This effect is minimized by reducing ℓ_D.

One of the limits on scaling is punch through voltage, which occurs when the field strength is too high, causing current to flow when the device is "turned off". Punch through voltage is a function of channel length (ℓ) and doping concentration (C_B), thus channel shortening can be compensated by increasing the doping



PERFORMANCE FACTORS

• HIGH DEVICE GAIN	GAIN $\alpha 1/(T_{OX}\ell)$
• LOW DIFFUSION CAPACITANCE	$C_P \alpha X_j$
• LOW MILLER CAPACITANCE	$C_m \alpha \ell_D$
• LOW BODY EFFECT	$\Delta V_T \alpha \sqrt{C_B}\, T_{OX}$

LIMITS

• PUNCH THROUGH VOLTAGE	$V_{PT} \alpha C_B \ell^2$
• THRESHOLD VOLTAGE	$V_T \alpha \sqrt{C_B}\, T_{OX}$

RESULT

- DECREASE ℓ, T_{OX}, X_j, ℓ_D
- INCREASE C_B

ℓ = CHANNEL LENGTH
T_{OX} = OXIDE THICKNESS
X_j = DIFFUSION DEPTH
ℓ_D = GATE OVERLAP
C_B = CONCENTRATION

Figure 1. HMOS Scaling

concentration. This has the additional advantage of balancing the threshold voltage which was decreased by scaling the oxide thickness for gain.

Comparison

Comparing scaling theory to HMOS II scaling in Table I, note that HMOS II agrees with scaling theory except for the supply voltage. It is left constant at +5V to maintain TTL compatibility. Had the voltage been scaled, the power would have been reduced by $1/K^3$ rather than $1/K$, but the device would not have been TTL compatible.

Table I. Scaling

	Theory	HMOS II
Dimensions	1/K	1/K
Substrate Doping	K	K
Voltage	1/K	1
Device Current	1/K	1
Capacitance A/T	1/K	1/K
Time Delay VC/I	1/K	1/K
Power Dissipation VI	$1/K^2$	1
Power Delay Product	$1/K^3$	1/K

THE DEVICE

The 2147H is TTL compatible, operates from a single +5 volt supply, and is easy to use.

Figure 2 shows the pin configuration and the logic symbol. The 2147H is compatible with the 2147 allowing easy system upgrade. Contained in an industry standard 18-pin dual in-line package the 2147H is organized as 4096 words of 1 bit. To access each of these words, twelve address lines are required. In addition, there are two control signals: $\overline{CS}$, which activates the RAM; and $\overline{WE}$,

which controls the write function. Separate data input and output are available. Logical operation of the 2147H is shown in the truth table. The output is in the high impedance or three-state mode unless the RAM is being read. Power consumption switches from standby to active under control of $\overline{CS}$.

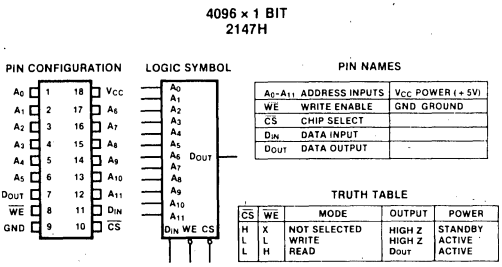


Figure 2. 2147H Logic Diagram

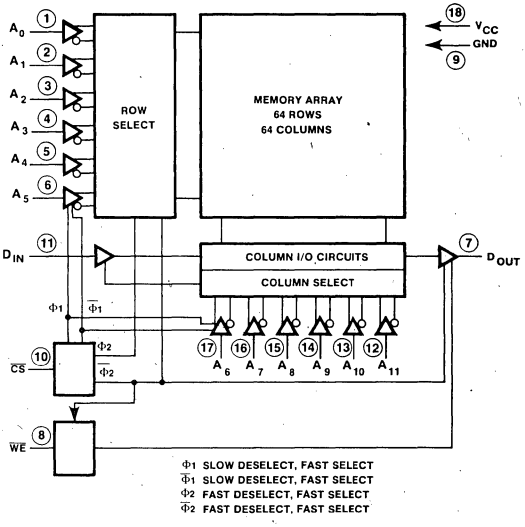


Figure 3. 2147H Block Diagram

Internal structure of the 2147H is shown in the block diagram of Figure 3. The major portions of the device are: addresses, control ($\overline{CS}$ and $\overline{WE}$), the memory array and a substrate bias generator, which is not shown.

The memory is organized into a two-dimensional array of 64 rows and 64 columns of memory cells. The lower-order six addresses decode one of 64 to select the row while the upper-order six addresses decode to select one column. The intersection of the selected row and the selected column locate the desired memory cell. Additional logic in the column selection circuit controls the flow of data to the array and as stated in the truth table, $\overline{WE}$ controls the output buffer.

As shown in Figure 4, the first three stages of the address buffer are designed with an additional transistor. In each stage, the lowest transistors are the active devices, the middle transistors are load devices, while the upper transistors, controlled by Φ_1, are the key to low standby power. Forming an AND function with the active devices, the upper transistors are turned off when the 2147H is not active, minimizing power consumption. Without them, at least one stage of these cascaded amplifiers would always be consuming power.

The signal Φ_1, and its inverse $\overline{\Phi}_1$, are generated from CS. They are part of an innovative design not found in the earlier 2147. Their function is to minimize the effects at short deselect times on the Chip Select access time, t_{ACS}.

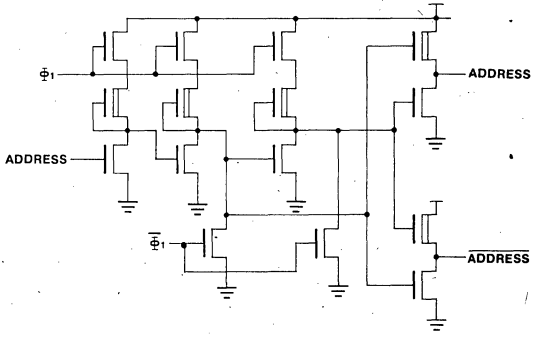


Figure 4. Address Buffer.

For both the 2147 and the 2147H, access is delayed until the address buffers are activated by chip selection. In the standard 2147, priming during deselection compensates for this delay by speeding up the access elsewhere in the circuitry. For short deselect times, however, full compensation does not occur because priming is incomplete. The result is a pushout in t_{ACS} for short deselect times.

In the 2147H, the address buffers are controlled by Φ_1, which is shaped as shown in Figure 5. Φ_1 activates rapidly for fast select time. However, Φ_1 deactivates slowly, keeping the address buffers active during short deselect times to speed access. As shown in Figure 6, this design innovation keeps t_{ACS} pushout to less than 1 ns.

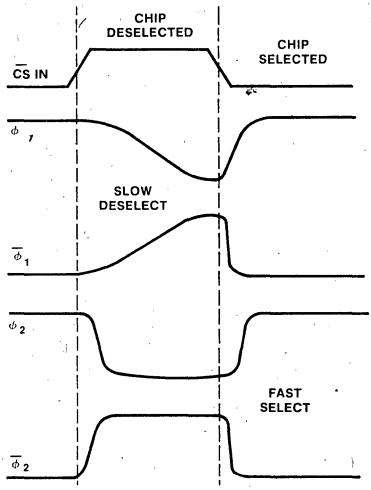


Figure 5. CS Buffer Signals

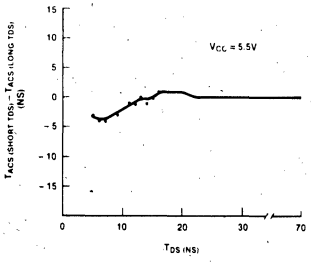


Figure 6. CS Access Vs. Deselect Time

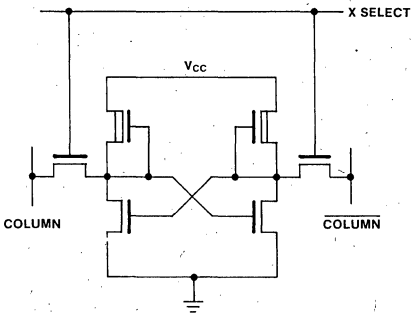


Figure 7. 2147H Memory Cell

Figure 7 shows the standard six-transistor cell. Configured as a bi-stable flip-flop, the memory cell uses two transistors for loads and two for active devices so that the data is stored twice as true and compliment. The two remaining transistors enable data onto the internal I/O bus. Unlike the periphery, the cell is not powered down during deselect time to sustain data indefinitely.

The 2147H has an internal bias generator. Bias voltage allows the use of high resistivity substrate by adjusting the threshold voltages. In addition, it reduces the effect of bulk silicon capacitance. As a result, performance is enhanced. Bias voltage is generated by capacitively coupling the output of a ring oscillator to a charge pump connected to the substrate. Internally generated bias permits the 2147H to operate from a single +5 volt supply, maintaining TTL compatibility.

2147H SUBSTRATE BIAS GENERATOR

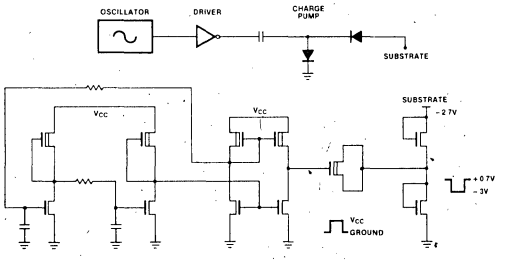


Figure 8. 2147H Substrate Bias Generator

DEVICE OPERATION

READ MODE

With power applied and $\overline{CS}$ at greater than 2V, the 2147H is in the standby mode, drawing less than 30mA. Activating $\overline{CS}$ begins access of the cell as defined by the state of the addresses. Data is transferred from the cell to the output buffer. Because the cell is static, the read operation is non-destructive. Device access and current are shown in Figure 9. Maximum access relative to the leading edge of $\overline{CS}$ is 35 ns for a 2147H-1. Without clocks, data is valid as long as address and control are maintained.

WRITE MODE

Data is modified when the write enable $\overline{WE}$ is activated during a cycle. At this time, data present at the input is duplicated in the cell specified by the address. Data is latched into the cell on the trailing edge of $\overline{WE}$, requiring that setup and hold times relative to this edge be maintained.

Two modes of operation are allowed in a write cycle, as shown in Figure 10. In the first mode, the write cycle is controlled by $\overline{WE}$, while in the other cycle, the cycle is controlled by $\overline{CS}$. In a $\overline{WE}$ controlled cycle, $\overline{CS}$ is held active while addresses change and the $\overline{WE}$ signal is pulsed to establish memory cycles. In the $\overline{CS}$ controlled cycle, $\overline{WE}$ is maintained active while addresses again change and $\overline{CS}$ changes state to define cycle length. This flexible operation eases the use and makes the 2147H applicable to a wide variety of system designs.

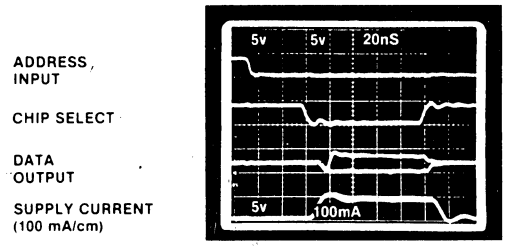
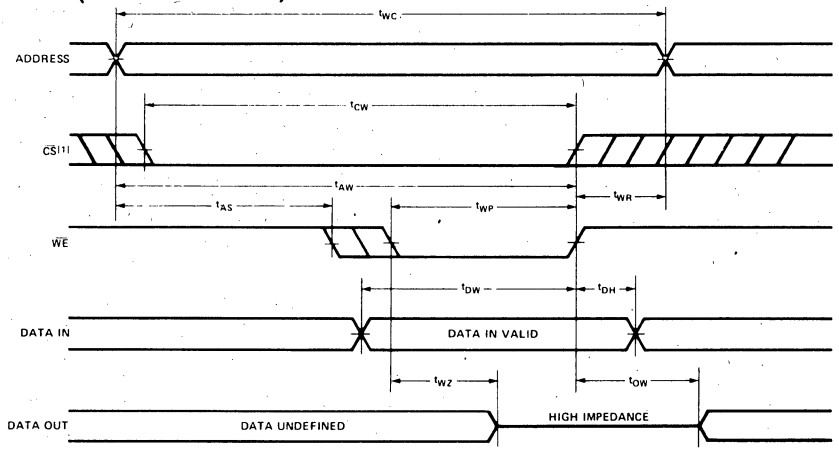


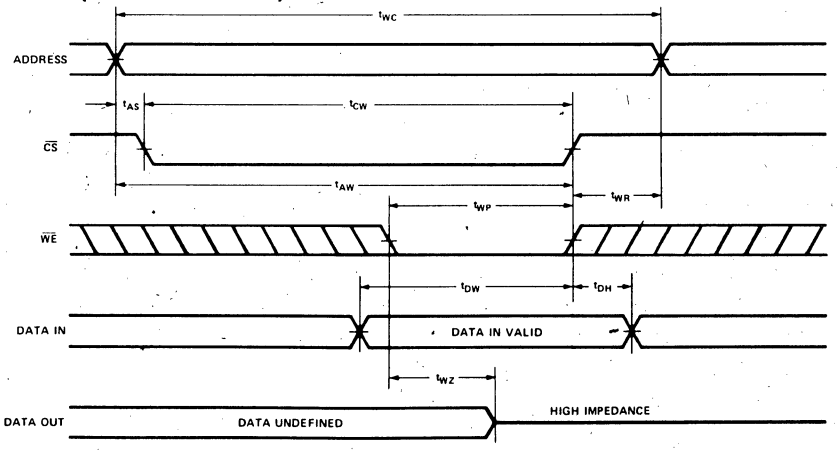
Figure 9. 2147H Access and Power Photo

WAVEFORMS

WRITE CYCLE #1 ($\overline{WE}$ CONTROLLED)



WRITE CYCLE #2 ($\overline{CS}$ CONTROLLED)



Note: 1. If $\overline{CS}$ goes high simultaneously with $\overline{WE}$ high, the output remains in a high impedance state.

Figure 10. Write Cycle Modes of Operation

EFFECT OF POWER DOWN AT THE SYSTEM LEVEL

Power consumed by a memory system is the product of the number of devices, the voltage applied, and the average current:

Equation 1

$$P = NVI_{AVE}$$

where:
P = Power
N = Number of devices
V = Voltage applied
I_{AVE} = Average current/device

Without power down, the average current is approximately the operating current. System power increases linearily with the number of devices. With power down, power consumption increases in proportion to the standby current with increasing number of memory devices. Curves in Figure 11 illustrate the difference which results from the majority of devices being in standby with a very small portion of the devices active or being accessed. For a system with power down, the average current of a device in the system is the sum of total active current and the total standby current divided by the number of devices in the system. For an X1 memory such as the 2147H, the number of active devices in most systems will be equal to the number of bits/word, m. Therefore, the number of devices in standby is the difference between N and M. I_{AVE} is expressed mathematically:

Equation 2

$$I_{AVE} = \frac{mI_{ACT} + (N-m)\, I_{SB}}{N}$$

where:
m = Number of active devices
I_{ACT} = Active current
I_{SB} = Standby current

The graph of Figure 12 shows the relation between average device current and memory size for automatic power down. For large memories the average device current approaches the standby current. Total system power usage, P, is calculated by substituting Equation 2 into Equation 1.

$$P = V[mI_{ACT} + (N-m)\, I_{SB}]$$

Comparison of power consumption of a system with and without power down illustrates the power savings. Assume a 64K by 18-bit memory constructed with 4KX1 devices. Active current of one device is 180mA and standby current is 30mA. Duty cycle is assumed to be 100% and voltage is 5 volts. The number of devices in the system is:

$$N = \frac{64K \text{ words} \times 18 \text{ bits/word}}{4K \text{ bit/device}}$$

$$N = 288 \text{ devices}$$

WITHOUT POWER DOWN:

$$P_{NPD} = 288 \text{ devices} \times 5 \text{ volts} \times 180 \text{ mA/device}$$
$$P_{NPD} = 259.2 \text{ watts}$$

WITH POWER DOWN:

With power down only 18 devices are active — 18 bits/word — and 270 are in standby.

P_{WPD} = 5 volts [18 devices (180mA/device) + 270 devices (30 mA/device)]

P_{WPD} = 56.7 watts

The system with power down devices uses only 22% of the power required by a non-powerdown memory system.

EFFECT OF POWER DOWN AT THE SYSTEM LEVEL

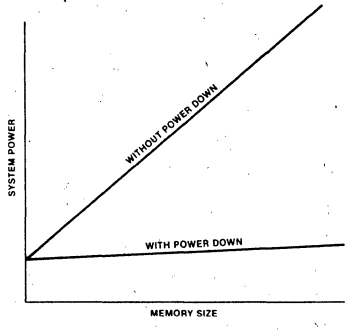


Figure 11. Effect of Power Down at the System

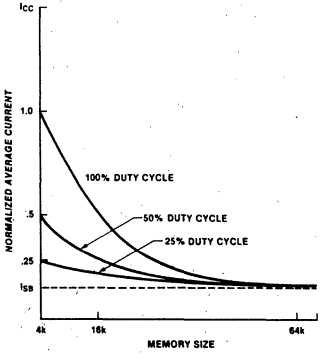


Figure 12. Average Current as a Function of Memory Size

POWER-ON

When power is applied, two events occur that must be considered: substrate bias start up and TTL instability. Without the bias generator functioning (V_{CC} less than 1.0 volts), the depletion mode transistors within the device draw larger than normal current flow. When the bias generator begins operation (V_{CC} greater than 1.0 volts), the threshold of these transistors is shifted, decreasing the current flow. The effect on the device power-on current is shown in Figure 13.

For V_{CC} values greater than 1.0 v., total device current is a function of both the substrate bias start-up characteristic and TTL stability. During power-on, the TTL circuits are attempting to operate under conditions which violate their specifications; consequently the $\overline{CS}$ signals can be indeterminent. One or several may be low, activating one or more banks of memory. The combined effects of this and the substrate bias start-up characteristic can exceed the power supply rating. The V-I characteristic of a power supply with fold back reduces the supply voltage in this situation, inhibiting circuit operation. In addition, the TTL drivers may not be able to supply the current to keep the $\overline{CS}$ signals deactivated.

One of several design techniques available to eliminate the power-on problem is power supply sequencing. Memory supply voltage and TTL supply voltage are separated, allowing the TTL supply to be activated first. When all the $\overline{CS}$ signals have stabilized at 2.0V or greater, the memory supply is activated. In this mode the memory power-on current follows the curve marked $\overline{CS}$ = V_{CC} in Figure 13.

If power sequencing is not practical, an equally effective method is to connect the $\overline{CS}$ signal to V_{CC} through a 1KΩ resistor. Although this does not guarantee a 2.0V $\overline{CS}$ input; emperical studies indicate that the effect is the same.

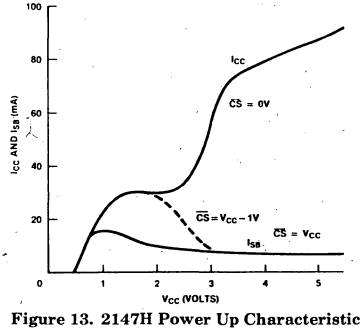


Figure 13. 2147H Power Up Characteristic

ARRAY CHARACTERISTICS

When two or more RAMs are combined, an array is formed. Arrays and their characteristics are controlled by the printed circuit card which is the next most important component after the memory device itself. In addition to physically locating the RAMs, the p.c. board must route power and signals to and from the RAMs.

GRIDDING

A power distribution network must provide required voltage, which from the 2147H data sheet is 5.0 volts ±10% to all the RAMs. A printed circuit trace, being an extremely low DC resistance, should easily route +5v DC to all devices. But as the RAMs are operating, micro circuits within the RAMs are switching micro currents on and off, creating high frequency current transients on the distribution network. Because the transients are high frequency, the network no longer appears as a "pure" low resistance element but as a transmission line. The RAMs and the lumped equivalent circuits of the transmission line are drawn in Figure 14. Each RAM is separated by a small section of transmission line both on the +voltage and the —voltage. Associated with the transmission lines is a voltage attenuation factor. In terms of AC circuits, the voltage across the inductor is the change in current — switching transient — multiplied by the inductance.

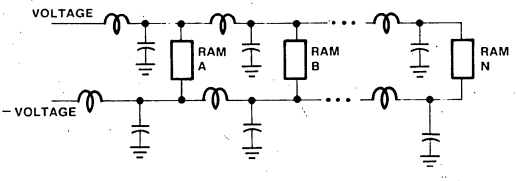


Figure 14. Equivalent Circuit for Distribution

Assuming all RAMs act similarly, the first inductor will see N current transients and the inductor at RAM B sees N-1 transients. The total differential is:

$$\Delta V = \sum_{n=1}^{N} n L \frac{di_n}{dt}$$

That voltage tolerance of ±10% could easily be exceeded with excursions of ±1 volt not uncommon. Measures must be taken to prevent this. The characteristic impedance of a transmission line is shown in Figure 15A.

Connecting two transmission lines in parallel will halve the characteristic impedance. The result is shown in Figure 15B.

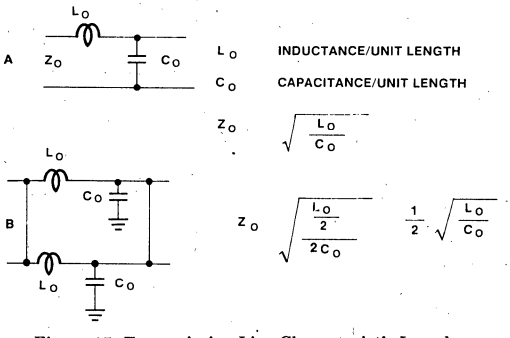


Figure 15. Transmission Line Characteristic Impedance

Paralleling N traces will reduce the impedance to Zo/N. Extrapolation of this concept to its limit will result in an infinite number of parallel traces such that they are physically touching, forming an extremely wide, low impedance trace, called a plane. Distribution of power (+ voltage) and ground (- voltage) via separate planes provides the best distribution.

P.C. boards with planes are manufactured as multi-layer boards sandwiching the power and ground planes internally. Characteristics of a multilayer board can be cost effectively approximated by gridding the power and ground distribution. Gridding surrounds each device with a ring of power and ground distribution forming many parallel paths with a corresponding reduction of impedance. Gridding is easily accomplished by placing horizontal traces of power (and ground) on one side of the pc board and vertical traces on the other, connected by plated through holes to form a grid.

Viewed from the top of the p.c. board, the gridding as in Figure 16 surrounds each device. Pseudo-gridding techniques such as serpentine or interdigitated distribution, as in Figure 17, are not effective because there are no parallel paths to minimize the impedance.

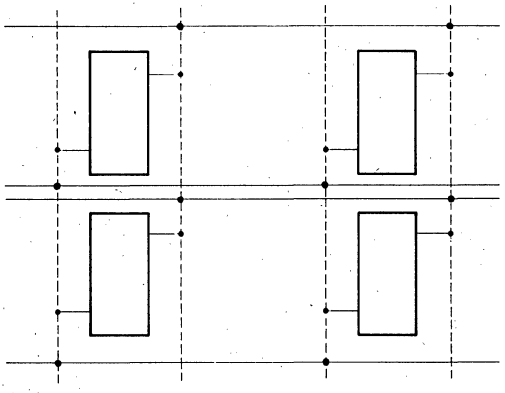


Figure 16. Gridding Plan

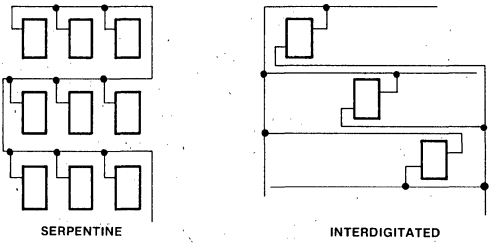


Figure 17. Pseudo-Gridding Techniques

DECOUPLING

One final aspect of power/ground distribution must be considered - decoupling.

Decoupling provides localized charge to minimize instantaneous voltage changes on the power grid due to current changes. These transient current changes are local and high frequency as devices are selected and deselected. Adequate decoupling for the 2147H is accomplished by placing a 0.1 μf ceramic capacitor at every other device as shown in Figure 18. Bulk decoupling is included on the board to filter low frequency noise in the system power distribution. One tantalum capacitor of 22 to 47 μf per 16 devices provides sufficient energy storage. By distributing these capacitors around the board several small currents exist rather than one large current flowing everywhere. Smaller voltage differentials - voltage is proportional to current - are experienced and the voltage remains in the specified operating range. Figure 19 demonstrates the difference with and without gridding.

TERMINATION

Similar reasoning is applied to the a.c. signals: address, control, and data. While they are not gridded or decoupled, they must be kept short and terminated. Similar to the power trace, the signal

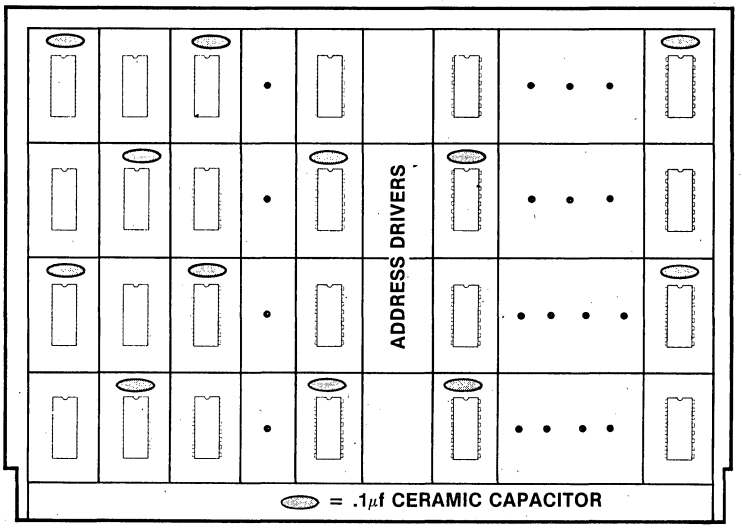
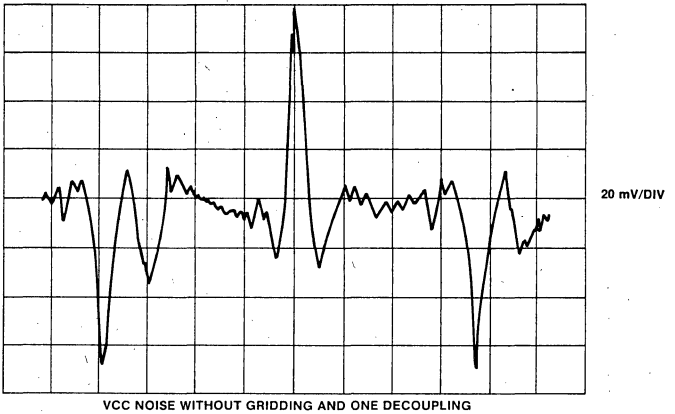
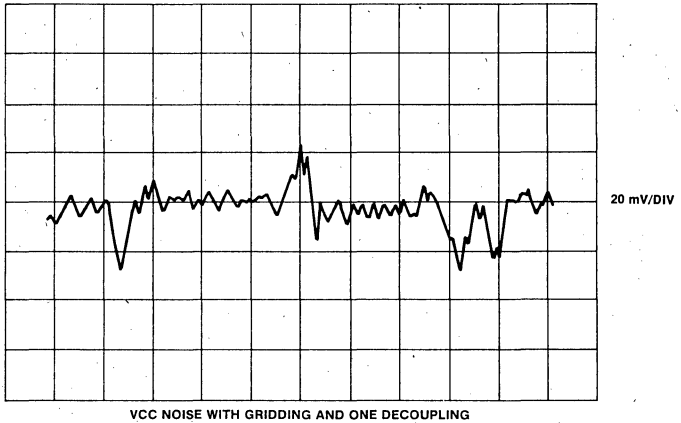


Figure 18. Decoupling



VCC NOISE WITHOUT GRIDDING AND ONE DECOUPLING CAPACITOR PER 4 RAMS



VCC NOISE WITH GRIDDING AND ONE DECOUPLING CAPACITOR PER 2 RAMS

Figure 19. V_{CC} Noise With & Without Gridding

trace will have transmission line characteristics. A simplified circuit is shown in Figure 20.

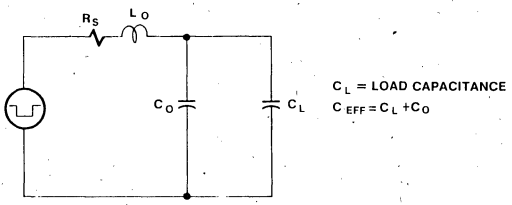


Figure 20. Signal Equivalent Circuit

MOS RAM input is essentially capacitive. Simplifying the capacitance and writing the differential equation.

$$\vartheta = L\frac{di}{dt} + \frac{1}{C}\int idt$$

The solution of this equation is:

$i = K_1 e^{-r_1 t} + K_2 e^{-r_2 t}$

where:

$$r_1 = \frac{R}{2L} + \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}}$$

$$r_2 = \frac{R}{2L} - \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}}$$

K_1 = constant

K_2 = constant

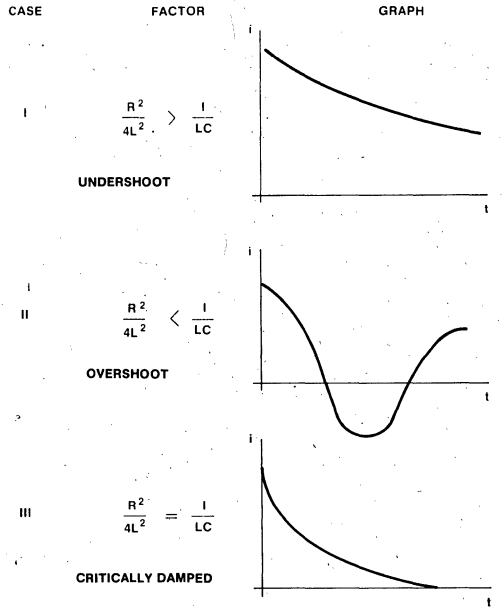


Figure 21. Three Cases of Equation Solution

Dependent on the values of R, L and C, there are three cases shown in Figure 21. In case I, rise and fall times are excessively long. In case III, the current smoothly and clearly changes, while in case II, the current overshoots and rings. If ringing is severe enough, the voltage can cross the threshold voltage of the device as in Figure 22.

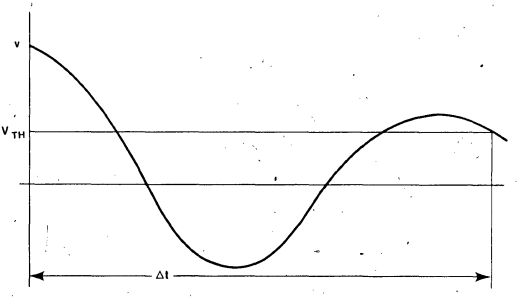


Figure 22. Access Push-Out Due to Ringing

Effective access is stretched out until the wave form settles. System access is the settling time (Δt) plus the specified device access. Case III is the ideal case but in reality a compromise between case I and case II is used because parameters vary in a production environment. Enough series resistance is inserted to prevent ringing but not enough to significantly slow down the access. A series resistance of 33Ω provides this compromise. The exact value is determined emperically but 33Ω is a good first approximation.

SERIES TERMINATION/ PARALLEL TERMINATION

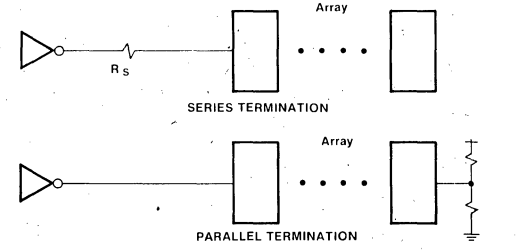


Figure 23. Series and Parallel Termination

Series termination uses one resistor and consumes little power. Current through the resistor creates a voltage differential shifting the levels of input voltage to the devices slightly. This shift is usually insignificant because the 2147H has an extremely high input impedance.

Termination could also be accomplished by a parallel termination as shown in Figure 23.

Parallel termination has the advantage of faster rise and fall times but the disadvantage of higher power consumption and increased board space usage.

SYSTEM DELAYS

RAMs are connected to the system through an interface, comprised of address, data and control signals. Inherent in the interface is propagation delay. Added to the RAM access time, propagation delay lengthens system access time and hence system cycle time. Expressed as an equation:

$$t_{sa} = t_{da} + t_{pd}$$

where: t_{sa} = system access time
t_{da} = device access time
t_{pd} = propagation delay

Device access is a fixed value, guaranteed by the data sheet. System efficiency then, is a function of system access and can be expressed as:

$$Eff = t_{da}/t_{sa}$$

where: Eff = System Efficiency

This can be reduced by substitution for tsa to:

$$Eff = 1/(1 + t_{pd}/t_{da})$$

System efficiency is maximized when propagation delay is minimized. With sub 100 ns access RAMs, efficiency can be reduced to 40-60% because delay through the signal paths is significant when compared to RAM access. Three factors contribute to the delay: logic delay, capacitive loading, and transit time.

LOGIC DELAY

The delay through a logic element is the time required for the output to switch with respect to the input. Actual delay times vary. Maximum TTL delays are specified in catalogs, while minimum delays are calculated as one-half of the typical specification. As an example, a gate with a typical delay of 6 ns has a minimum delay of 3 ns.

A signal propagating through two logically identical paths but constructed from different integrated circuits will have two different propagation times. For example, in Figure 24A one path has minimum delays while the other has maximum delays. Path A-B has a delay of 3.5 ns while A-B^1 has a delay of 11 ns. The time difference between these two signals is skew, which will be important later in the system design. Figure 24B shows skew values for several TTL devices.

CAPACITIVE LOADING

Delay time is also affected by the capacitive load on the device. Typical delay as a function of capacitive load is shown in Figure 25. TTL data sheets specify the delay for a particular capacitive load

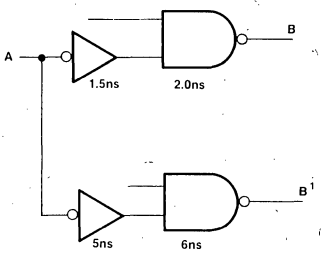


Figure 24A.

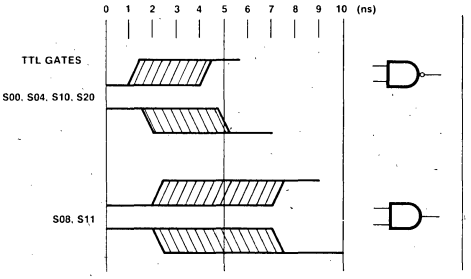


Figure 24B. Skew

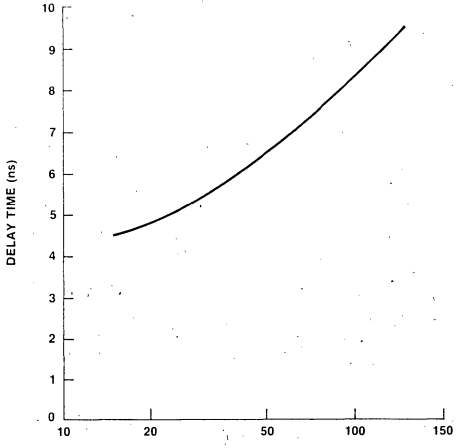


Figure 25. Capacitive Loading

(typically 15pF or 50 pF). Loads greater than specified will slow the device; similarly, loads less than specified will speed up the device.

A value of 0.05 ns/ft is a linear approximation of the function in Figure 25 and is used in the calculations. Loading effect is calculated by subtracting the actual load from the specified load. This difference is multiplied by 0.05 ns/pF and the result algebraically subtracted from the specified delay. As an example, a device has a 4 ns delay driving 50 pF, but the actual load is 25 pF. Then,

50 pF specified
-25 pF actual
25 pF difference

25 pF × 0.05 ns/pF = 1.25 ns

4 ns specified
-1.25ns difference
2.75 ns actual delay

A device specified at 4 ns while driving 50 pF will have a delay of only 2.75 ns when driving 25 pF. Conversely, the same device driving 75 pF would have a propagation time of 5.25 ns.

TRANSIT TIME

Signal transit time, the time required for the signal to travel down the P.C. trace, must also be considered. As was shown in Figure 19, these traces are transmission lines. Classical transmission line theory can be used to calculate the delay:

$$t_p = \sqrt{LC}$$

where: t_p = Travel Time
L = Inductance/unit length of trace
C = Capacitance/unit length of trace

The capacitance term in the equation is modified to include the sum of the trace capacitance and the device capacitance. This equation approximates in the worst case direction; a signal will never "see" all the load capacitance simultaneously, it is distributed along the trace at the devices.

Substituting into the equation:

$$tp^1 = \sqrt{L(C + C_L)}$$

where: tp^1 = Modified delay
C_L = Load capacitance

Algebraically:

$$tp^1 = \sqrt{LC(1 + C_L/C)}$$

$$tp^1 = \sqrt{LC}\ \sqrt{1 + CL/C}$$

and

$$tp^1 = tp\ \sqrt{1 + CL/C}$$

Emperically, tp is 1.8 ns/ft for G-10 epoxy and C is 1.5 pF/in. For a 5-in. trace and a 40 pF load, the delay is calculated to be 4.5 ns. Because this is worst case, an approximated 2 ns/ft can be used. In the following sections, however, the equation will be used. Total delay is the summation of all the delays. Adding the device access, TTL delays and the trace delays result in the system access.

BOARD LAYOUT

The preceding section discussed the effects of trace length and capacitive loading. Proper board layout minimizes these effects.

As shown in Figure 26, address and control lines are split into a right- and left-hand configuration with these signals driving horizontally. This configuration minimizes propagation delay. Splitting the data lines is not necessary, as the data loads are not as great nor are their traces as long as address and control lines. Control and timing fills the remaining space.

Two benefits are derived from this layout. First,

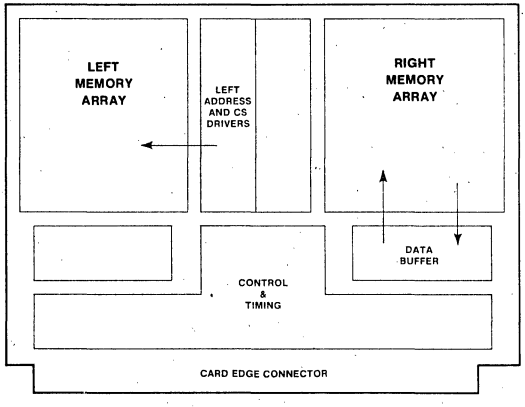


Figure 26. Board Layout

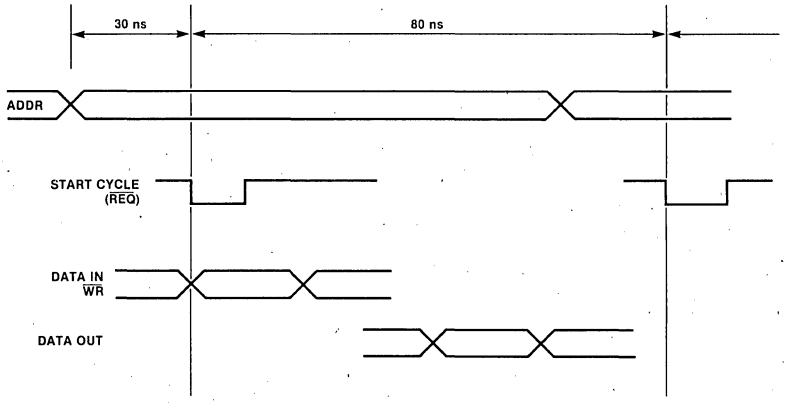


Figure 27. System Timing

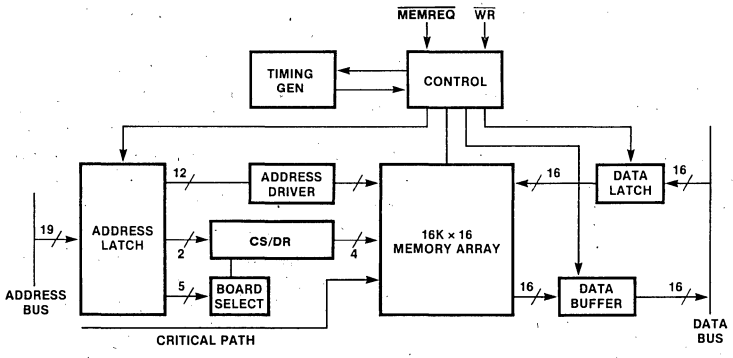


Figure 28. System Block Diagram

the address and control lines are perpendicular to the data lines which minimizes crosstalk. Second, troubleshooting is simplified. A failing row of devices indicates a defective address or control driver; whereas a failing column indicates a faulty data driver.

SYSTEM DESIGN

Using previously discussed rules and guidelines, the design of a typical high speed memory will be reviewed to illustrate these techniques. Configuration of the system is a series of identical memory cards containing 16K words of 16 bits. Timing and control logic is contained on each board. System timing requires an 80 ns cycle as shown in Figure 27. Cycle operation begins when data and control signals arrive at the board. In this design, addresses are shifted 30 ns to be valid before the start of the cycle so that address, data, and control arrive at the memory device at the same time for maximum performance. Data and control signals are coincident with the start of the cycle. Access is not yet specified because it is affected by device access and the unknown propagation delay. Access will be determined in the design.

Figure 28 illustrates the elements of the system in block diagram form. Addresses are buffered and latched at the input to the printed circuit card. Once through the latch, the addresses split to perform three functions: board selection, chip select ($\overline{CS}$) generation, and RAM addressing. Highest order addresses decode the board select, which enables all of the board logic including $\overline{CS}$.

Next higher order addresses decode $\overline{CS}$, while the lowest order addresses select the individual RAM cell. Data enters the board from the bidirectional bus through a buffer/latch, while output data returns to the bidirectional bus via buffers. Only two control signals — cycle request ($\overline{MEMREQ}$) and write ($\overline{WR}$) control the activity on the board.

Figure 29 illustrates the levels of the delay in the

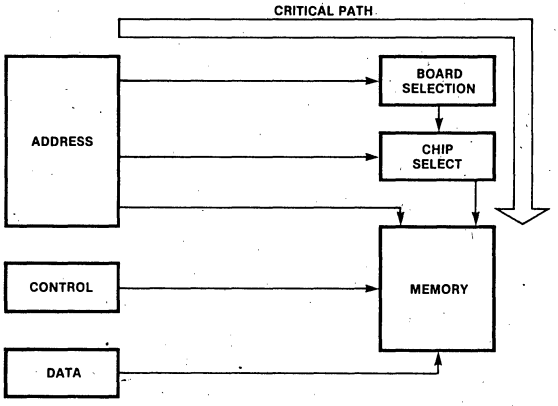


Figure 29. Worst Case Delay Path

system. Data and control have only one level. But examine the address path, it has three levels. Addresses are decoded to activate the logic on the board, select the row of RAM to be accessed and finally locate the specific memory cell. $\overline{CS}$ is in this address path and is crucial for access; without it RAM access cannot begin. But this path has the most levels of decoding with associated propagation delays. Consequently, the address path to $\overline{CS}$ is the critical path and has the greatest effect on system delay and hence must be minimized.

Examination of the system begins with the $\overline{CS}$ portion of the critical path, followed by addresses, data path, and finally timing and control.

CRITICAL PATH

Analysis of the critical path begins with the address latch. The first decision to be made is to the latch type. Latches can be divided into two types: clocked and flow-through. Clocked latches capture the data on the leading or trailing edge of the clock. Associated with the clock is data set-up or hold-time that must be included in the delay time. Accuracy of the clock affects the transit time of the signal because any skew in the clock adds to the delay time. As an example, a typical 74S173 latch has a data set-up time of 5 ns and a maximum propagation delay time from the clock of 17 ns. Total delay time is 22 ns, excluding any clock skew.

Flow-through latches have an enable rather than clock. The enable opens the address window and allows addresses to pass independent of any clock. Delay time is measured from the signal rather than a clock. The Intel® 3404 is a high speed, 6-bit latch operating in a flow-through mode with 12 ns delay. This is acceptable but a faster latch can be fashioned using a 2-to-1 line multiplexer, either a 74S157 or a 74S158. The slower of the two is the 74S157 with 7.5 ns delay. Although the 74S158 is faster with 6 ns delay, it requires an extra inverter in the feedback path as shown in Figure 30. Between the 74S157 and the 74S158 latches, the trade off is speed against board space and power. Individual designers will choose to optimize their designs.

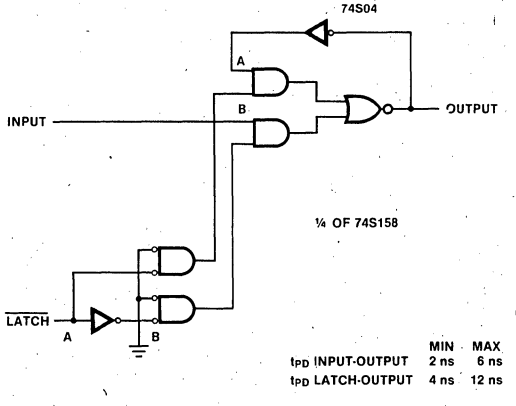


Figure 30. Fast Latch

In either case, care must be exercised in constructing the latch. Output data must be fed back to the input having the shortest internal path — the A input. If the latch is constructed with the output strapped to the B input, the input could be deselected and the feedback loop not yet selected because of the delay through the internal inverter. In this situation data would be lost. Additional delay through the external inverter (74S04) aids in preventing data loss. Inverting addresses has no system effect — except that it's faster than the non-inverting latch. During a write cycle, data will be stored at the compliment of the system address. When this data is to be retrieved, the same address will be complimented, fetching the correct word.

The remaining elements in the critical path to be designed are board selection and $\overline{CS}$ decoding. To minimize the $\overline{CS}$, decode path, the easiest method is to work backwards from $\overline{CS}$. In this manner input signals to a stage are determined and the output from the preceding stage is defined. This saves inserting an inverter at the cost of 5 ns to generate the proper input to a stage.

Starting with the $\overline{CS}$ driver, the design analyzes several approaches to select the fastest one. With four rows of devices, there are four $\overline{CS}$ signals to be generated. A 2-to-4 line decoder like the 74S138 is a possible solution. It is compact, but has two detriments: long propagation delay and insufficient drive capability. Propagation delay from enable is 11 ns. Enable is driven by board selection which arrives later than the binary inputs. Splitting the RAMs into two 4×8 arrays eases the drive requirement but the demultiplexer must still drive eight devices at 5 pF each — or 40 pF total — which adds 1.75 ns to the delay. More importantly, signal drive is required to switch cleanly and maintain levels in spite of crosstalk and reflections. A 74S240 buffer will solve this but in the process consumes an additional 9 ns.

A second and preferred approach is to use a discrete decoder to decode and drive the $\overline{CS}$ signals. Four input NAND buffers — 74S40 — fulfill this function. Addresses A_{12} and A_{13} are inverted via 74S04, providing true and compliment signals to the buffer for decoding. As shown in Figure 31, the delay is 11.5 ns. Propagation delay for the 74S40 is specified into a 50 pF load, eliminating the additional loading delay. Left and right drivers — CSXL and CSXR — are in the same package to minimize skew between left and right bytes of data. All of the decoders are enabled by Board Select to prevent rows of devices on several boards from being simultaneously active. Board Select is a true input, defining the output from the Board Select decoder.

In the Board Select decoder, the high order addresses are matched to hard-wired logic levels generated with switches for flexibility. Changing a switch setting shifts the 16K range of the board. Comparison of the switch setting and the address can be accomplished with an exclusive-OR, a 74S86. NANDing all the exclusive-OR outputs will generate a Board Select signal. Unfortunately, this signal is active-low, requiring an additional inverter as in Figure 32A, and it also consumes 22.5 ns to decode. An MSI solution to board selection is a 4-bit comparator — 74S85 — which

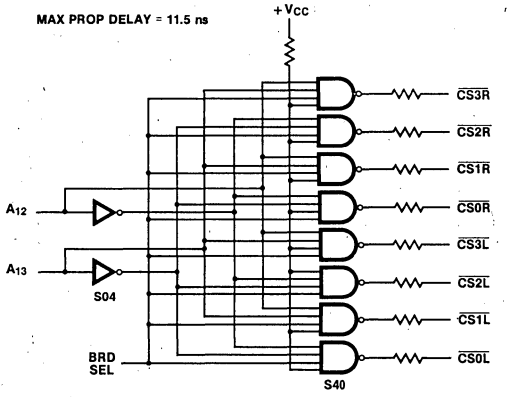


Figure 31. $\overline{CS}$ Decode

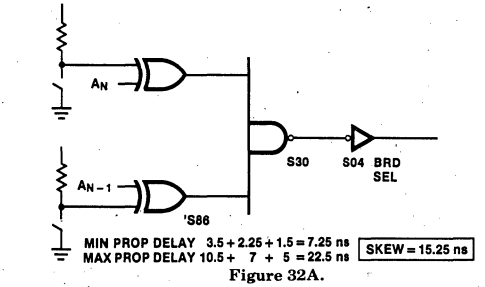


Figure 32A.

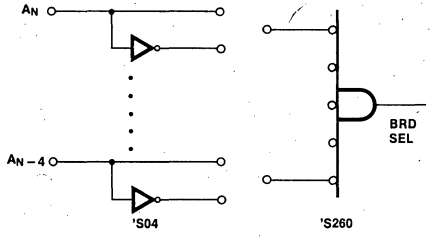


Figure 32B. Board Select

consumes less board area and propagation delay is improved at 16.5 ns.

The best solution is attained by inverting the high order addresses to generate true and compliment signals. the appropriate signal is connected into a 74S260, 5-input NOR. With an active-high output, maximum delay is 11 ns as in Figure 32B.

Critical path timing is the sum of the latch, Board Select, and $\overline{CS}$ delay times. In this example, latch delay is 6 ns, Board Select is 11 ns and $\overline{CS}$ decode is 11.5 ns for a total of 28.5 ns. One additional delay — trace delay — must be included for a complete solution. Each 74S40 drives eight MOS inputs having 5 pF/device for a load of 40 pF. Trace capacitance is calculated on 5 in. of trace. At 1.5 pF/in., trace capacitance is 7.5 pF. Trace delay calculated from equation 3 is 1.9 ns.

$$tp^1 = \frac{1.8\text{ ns}}{\text{ft}} \times \frac{5\text{ in.}}{12\text{ in./ft}} \sqrt{1 + \frac{40\text{ pF}}{7.5\text{ pF}}}$$

$$tp^1 = 1.9\text{ ns}$$

Total worst case maximum critical path delay has been calculated to be 30.4 ns (28.5 ns + 1.9 ns). With the addresses shifted in time by an amount equal to the worst case delay, device and system cycle start are coincident. Start of system access and device access differ only 0.4 ns when the addresses are shifted 30 ns. From the system cycle start, access is stretched by 0.4 ns as shown in Figure 33. Thus, with a 35 ns 2147H-1, data is valid at the output of the device 35.4 ns after the start of the cycle.

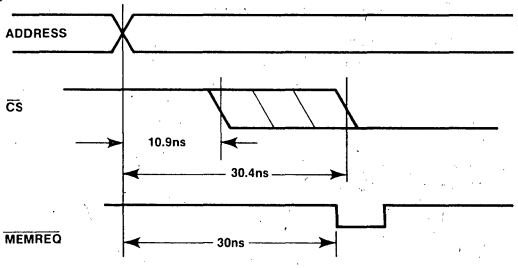


Figure 33. $\overline{CS}$ Decode Time

The minimum delay also must be calculated. With addresses valid prior to the start of the cycle, $\overline{CS}$ decoding can start in the previous cycle. If it occurs too soon, the previous cycle will not be properly completed. Minimum delay time is the sum of the minimum propagation delays plus capacitive loading delay plus trace delay. Capacitive loading delay is less than 0.4 ns and ignored. Minimum delay through the TTL is 9 ns, and added to trace delay results in a total of 10.9 ns.

From address change, the maximum delay in the critical path is 30.4 ns while the minimum is 10.9 ns. The difference between these two times is skew and will be important in later calculations.

ADDRESSES

Lower order addresses (A_0-A_{11}) arrive at the devices earlier than $\overline{CS}$ because they are not decoded. Consequently, the address drivers do not have a critical speed requirement. Once through the 6 ns latch, addresses have 24 ns to arrive at the devices.

While speed is not the primary prerequisite, drive capability is. Address drivers are located in the center of the board, dividing the array into two sections of 32 devices each. For the moment, assume one driver drives 32 devices as in Figure 34A. Each device is rated at 5 pF/input, resulting in a load of 160 pF. In addition, there are four 5-in. traces — one for each row. twenty inches of trace equates to 30 pF. Total capacitive load is 190 pF. A 74S04 is specified at 5 ns delay into 15 pF. The increased capacitive load is 175 pF, which at 0.05 ns/pF increases the delay by 8.75 ns. Under these conditions the worst cast driver relay is 5 ns plus 8.75 ns, totalling 13.75 ns. It is 10 ns earlier than the 24 ns available.

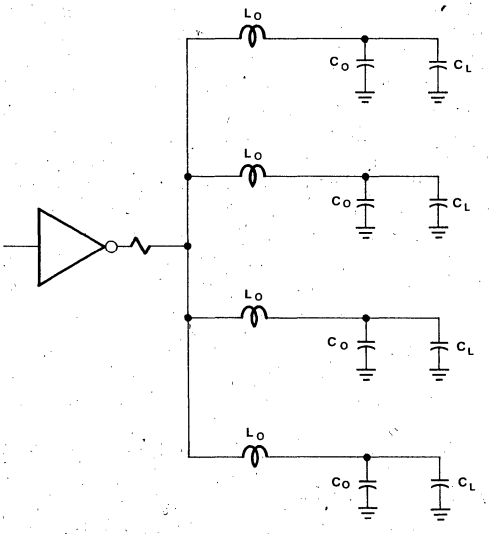


Figure 34A. Address Driver

The first impression is that this is sufficient, but the effect of crosstalk must be considered. For example, as shown in Figure 35, each trace has inductance, and parallel traces take on the

characteristics of transformers. When a signal switches from a one level to a zero level, its driver can sink 20 mA, inducing a transient in an adjacent trace. If the adjacent signal is switching to a one level, only 400 μA of a source current from the driver is available. The induced current will generate a negative spike, driving the signal at a one leval negative. Additional time of 10 to 15 ns is required to recover and re-establish a stable one level. This may prevent stable address at the start of the cycle. Recall:

$$i = C\frac{dv}{dt} \quad \text{or} \quad dt = C\frac{dv}{i}$$

where: i = instantaneous current
C = capacitance

$$\frac{dv}{dt} = \text{voltage time rate of change}$$

The term dv/dt can be maximized by increasing i or decreasing C. Current can be doubled by using a driver like a 74S240, but it draws 150mA supply current. In a large system the increased power is a disadvantage because it requires a larger power supply and additional cooling.

A better alternative is to reduce the capacitance, which results in a corresponding increase in dv/dt for quick recovery. Splitting the loads to 16 devices reduces the capacitance and allows a low power driver, like a 74S04, to be used, as in Figure 34B. This has the double effect of decreased propagation delay and providing sharp rise and fall times.

Now, there are only 10 in. of trace or 15 pF load and 16 devices, representing 80 pF for a total of 95 pF. Again, the S04 delay is 5 ns into 15 pF, but the stretched delay due to 80 pF is only 4.0 ns for a total of 9.0 ns. Stable addresses are guaranteed at the start of the cycle.

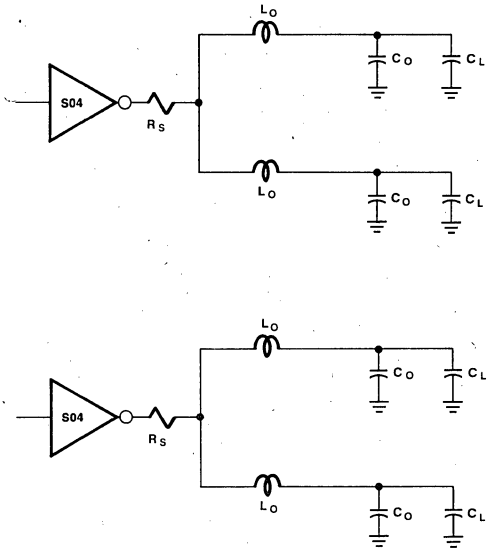


Figure 34B. Address Drivers

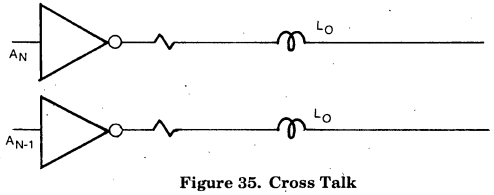


Figure 35. Cross Talk

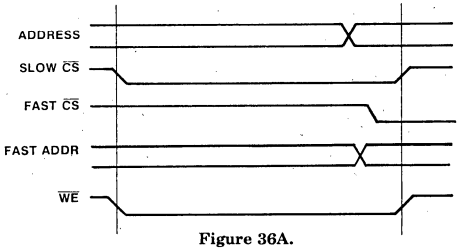


Figure 36A.

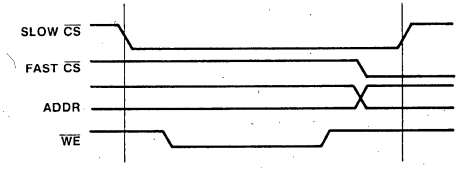


Figure 36B. Race Condition Between Address and WE

DATA PATH

Next in line for analysis is the data path. Reference to the system block diagram shows that the data is latched into the board on a write cycle and buffered out during a read cycle. Data latches are constructed from 74S158 quad two-input multiplexers. Because the data bus is bidirectional, 74S240 three-state drivers are used for output buffers.

All that remains to complete the board access computation is the calculation of the output propagation delay. Output delay of the active RAM is caused by the capacitance loading of its own output plus the three idle RAMs, the input capacitance of the 74S240 bus driver and trace capacitance. Output capacitance of the 2147Hs is 6 pF/device for a subtotal of 24 pF; input capacitance of the 74S240 is 3 pF and trace capacitance of a 5-in. trace is 7.5 pF. total load

capacitance is 34.5 pF, and access time of the 2147H is specified driving a 30 pF load. Calculated loading is close enough to the specified loading to eliminate any significant effect on the access calculations. Had there been a difference, the effect would have been included in the calculation. As previously calculated, transit time of the trace is 1.6 ns. Adding this to the 7 ns delay through the 74S240 bus driver results in an 8.6 ns output propagation delay from the RAM output to the bus.

Total access is 35.4 ns plus 8.6 ns output delay for a total access of 44 ns. The efficiency of this system is:

$$\text{Eff} = \frac{35}{44} \text{or } 80\%$$

TIMING AND CONTROL

Timing and control gating regulates activity on the board to guarantee operation in an orderly fashion. This gating latches addresses, controls the write pulse width and enables the three-state bus drivers. In addition, accurately generated timing compensates for skew effects.

In anticipation of the next cycle, the latch must be opened for the new address. When the current cycle has completed 50 ns, the latches are again opened. The next cycle might not begin 30 ns after the latch is opened because the system may skip one or more memory cycles. Therefore, a signal from the next active cycle must close the latch. In operation, a buffered Memory Request signal latches the addresses.

The write pulse is controlled to guarantee set-up and hold times for data and address and to prevent an overlap of $\overline{CS}$ and write enable from different cycles. To understand the consequences, consider the following example.

Assume two memory banks, one has a minimum $\overline{CS}$ and the other has a maximum delay path in $\overline{CS}$, and both have a minimum address delay. Assume that $\overline{WE}$ is a level generated from a write command as shown in Figure 36A. The operation under examination is a write cycle into the bank with fast $\overline{CS}$ followed by a read cycle into the bank with slow $\overline{CS}$.

Both the write cycle and the read cycle have device specification violations. In the write cycle, the addresses change prior to $\overline{CS}$ and $\overline{WE}$ becoming inactive; that new address location may be written into. In the read cycle, the address change is correct but $\overline{WE}$ is still active and the fast $\overline{CS}$ begins too soon, performing a non-existent write cycle. Clearly, controlling the width of $\overline{WE}$ will solve the problems.

Figure 36B shows the proper operation controlled with timing.

Finally, the data output buffers, controlled by timing signals, are enabled only during a read cycle while the board is selected preventing bus contention with two or more boards in the system. More importantly, timing disables the output prior to the start of the next cycle, allowing input data to be stabilized on the bidirectional data bus in preparation for a write cycle.

TIMING GENERATION

Having discussed the philosophy of timing and control, we can now focus on the specifics of address latching, write pulse generation and output-enable timing. To perform these functions timing can be generated from one of three sources: clock and shift register, monostable multivibrator, or delay line.

CLOCKED SHIFT REGISTER

A clocked shift register circuit is shown in Figure 37 consisting of a D-type flip flop and an 8-bit shift register.

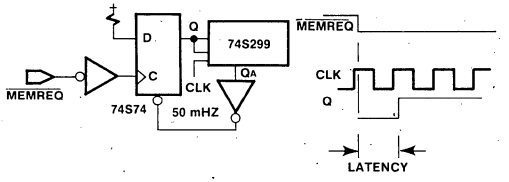


Figure 37. D Flip-Flop and Shift Register

On the leading edge of $\overline{MEMREQ}$, the Q output of the D flip flop is clocked to a one state, enabling a "one" to be propagated through the shift register. The one is clocked into the first stage of the shift register on the first clock edge after the A and B inputs are "ones". After the clock, the output Q_A goes true which subsequently clears the D flip flop, clocking zeros into the register to create a pulse one clock period wide.

The accuracy and repeatability depends primarily on the accuracy and stability of the clock. Crystal clocks can be built with +0.005% tolerance and less than a 1% variation due to temperature.

An inherent difficulty is the synchronization of Memory Request and the clock. At times there will be a latency of one clock cycle between Memory Request and the actual start of the cycle when Memory Request becomes active just after the clock edge. Assuming an 80 ns cycle and 20 ns clock, the latency can be 20 ns or 25% of a cycle stretching both access and cycle accordingly. A second difficulty of this circuit is caused by the asynchronous nature of the clock and the Memory Request. The request becomes active just prior to

the clock and the set-up time of the latch is violated, the output Q_A "hangs" in a quasi-digital state and could double or produce an invalid pulse width; this and the latency hinder effective use in high speed design.

MONOSTABLE MULTIVIBRATOR

The second possible timing generator is a series of monostable multivibrators, using a device such as the AMD Am 26S02 multivibrator. It has a maximum delay from input to output of 20 ns and an approximate minimum of 6 ns. However, with a delay of 20 ns, the monostable multivibrator offers no advantage over the clocked generator. Having a minimum pulse width of 28 ns, the one-shot offers no improvement over the 50 MHz clock, but in fact the performance is worse because it is more temperature and voltage sensitive. The pulse width is dependent on the RC network composed of resistors and capacitors that are temperature sensitive. Consequently, repeatability leaves something to be desired.

DELAY LINE

The third and best choice is a delay line. This design uses STTLDM-406 delay lines from EC[2] with tapped outputs at 5 ns increments. In operation, Memory Request activates an R-S flip flop fabricated from cross coupled NAND gates. The output of this circuit starts the memory cycle. Consequently, the cycle starts 5 ns after Memory Request compared to 20 ns for the other two timing generators. The leading edge travels down the delay lines. When the edge reaches the 25 ns tap, the output is inverted and fed back to the R input of the R-S flip flop, shaping the pulse to width to 25 ns. Twenty-five nanoseconds was chosen to match as close as possible the write pulse width. A 25 ns pulse limits the Memory Request signal width to less than 25 ns to insure proper operation. Otherwise, the R-S flip flop will not clear until Memory Request returns to a one level. As the pulse travels down the delay lines, it acquires additional skew of ±1 ns per delay line package for a total of 6 ns overall. Figure 38 shows several timing pulses and the uncertainty of each edge calculated by worst case timing analysis. The remaining problem is selection of timing edges to operate the device. Now that the timing chain is completely defined, specific details of the address latch, write pulse and output enable can be completed.

ADDRESS LATCH TIMING

An R-S flip flop activated by MEMREQ latches the addresses. A second signal which we will now calculate is used to open the latch. This signal has two boundaries. If the latch opens too late, the access of the cycle will be extended; if it opens too soon, the current cycle will be aborted. Skew through the R-S flip flop is 1.75 ns to 5.5 ns and skew in the latch from enable to output is 4 ns to 12 ns for a total skew of 6 to 17.5 ns. With this skew added to the 30 ns address set-up time, the latch opening signal must be valid at 36 ns best case or

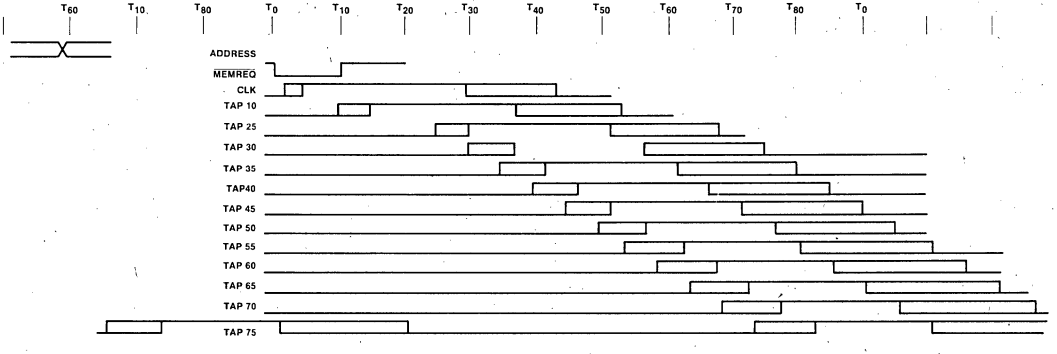


Figure 38. Timing Chain

47.5 ns worst case prior to the start of the memory cycle. Each cycle is 80 ns long, therefore, the latch opening signal must begin 44 ns or 32.5 ns, respectively, in the preceding cycle. From the delay line timing diagram, T35 will satisfy the worst case requirements for opening the latch and T 25 best case. In production, each board is tuned by selecting T25, T30, or T35 to open the latch, guaranteeing it opens between 35 and 30 ns prior to the start of the cycle.

WRITE PULSE TIMING

The next timing to be calculated is the write pulse. Figure 39 shows the three parameters which define the write pulse timing: data set-up time, write pulse width and write recovery time. Data set-up is assured by having data valid through the entire cycle.

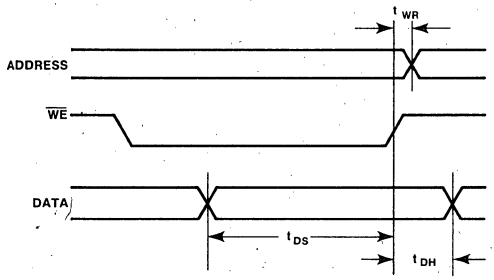


Figure 39. WE Constraints

Placement of $\overline{WE}$ in the cycle is controlled by address change to comply with t_{WR}. From previous calculations the earliest addresses can change is 50 ns, which defines the end of the $\overline{WE}$ signal. Our calculations begin at the device and work back to the timing edge. Eight devices constitute a 40 pF load and a 74S40 is specified for a 50 pF load, reducing delay by 0.5 ns when driving 40 pF. Trace delay and 74S40 delay is 3.5 to 8 ns. Subtracting 8 ns from 50 ns sets the termination of the write timing edge at 42 ns. Using the inversion of T25 will end the write pulse at 43 ns with 7 ns to spare.

Data set-up time is guaranteed because data is valid 6 ns (the worst case delay through the latch) after the start of $\overline{MEMREQ}$.

OUTPUT ENABLE TIMING

There is a 5.5 ns delay through the address driver providing minimum device cycle of 50 ns. As a result the earliest data can disappear from the bus is at 54 ns because of delay through the output circuit. To select the timing tap for the output enable, the skew of the enable circuit is subtracted from the system access time.

Subtracting the 28 ns skew of the buffer enable circuit from the 44 ns access time of the system shows that the latest the timing edge can occur is 16 ns, which is satisfied by edge T10. The trailing edge, however, ends at 37 ns and with minimum propagation delays the bus would become three-stated at 44 ns, coincident with data becoming valid. ORing T20 with T10 will guarantee the output is valid until 54 ns, minimum. Selecting a timing gap between T35 and T50, depending on the propagation delay in the enable circuit, disables the output at 70 ns, allowing input data to be valid for 10 ns prior to start of cycle. The complete schematic is shown in Figure 40.

SUMMARY

The 2147H is an easy-to-use, high speed RAM. The problems in a memory system design are the result of inherent limitations in interfacing. Largest of these is skew, which the designer must strive to minimize. In this example, skew consumed 45 ns of an 80 ns cycle while device access time was extended by only 10 ns, resulting in an 80% efficiency.

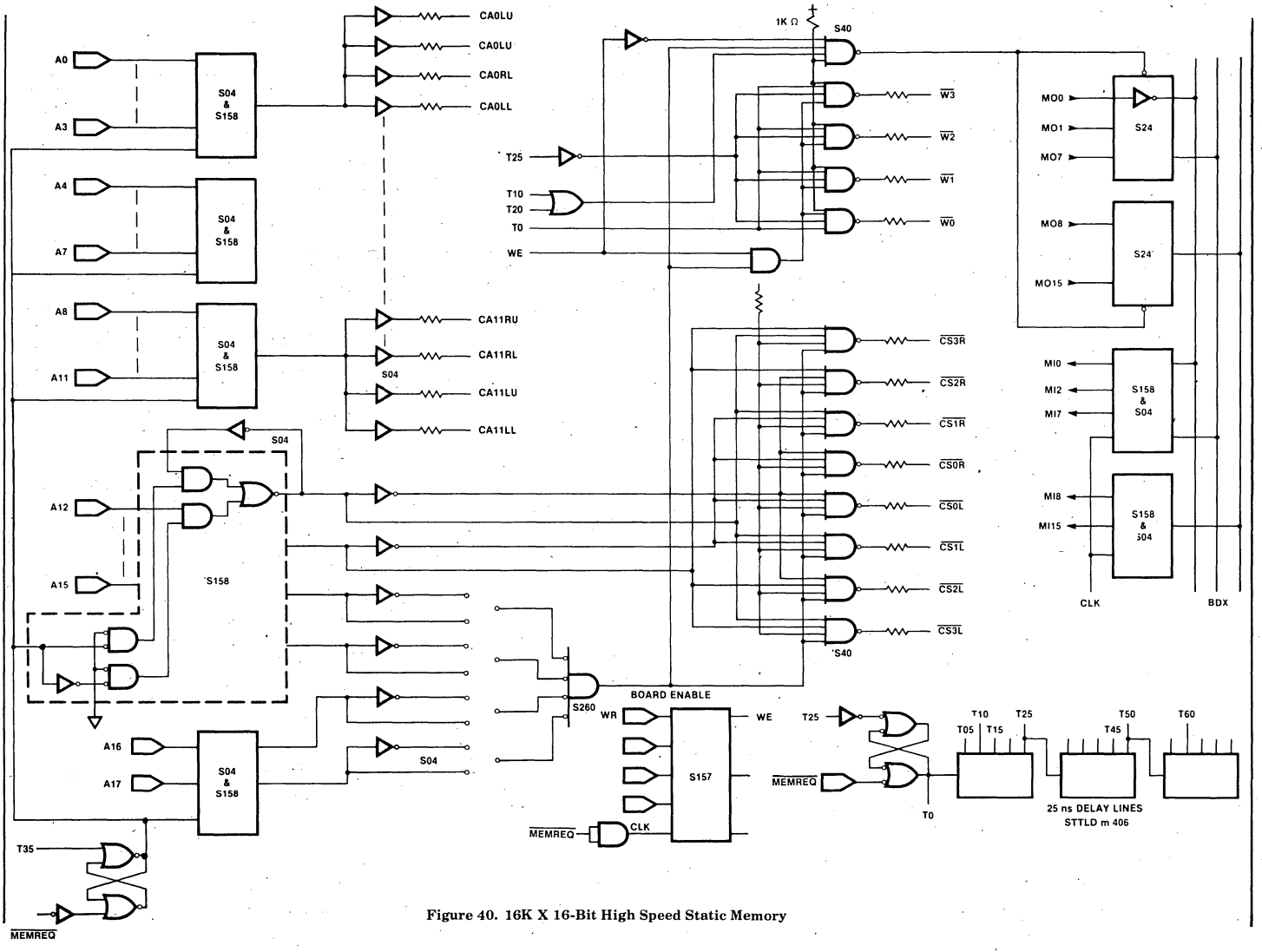


Figure 40. 16K X 16-Bit High Speed Static Memory